Application for United States Letters Patent

for

SRAM with Forward Body Biasing to Improve Read Cell Stability

by

Muhammad M. Khellah
Dinesh Somasekhar
Yibin Ye
Ali R. Farhang
Gunjan H. Pandya
Vivek K. De

Prepared by:

Seth Z. Kalson
Intel Corporation
2200 Mission College Blvd.
Santa Clara, CA 95052-8119

"Express Mail" mailing label number: EV409361719US
Date of Deposit: March 26, 2004
I hereby certify that I am causing this paper or fee to be deposited with the United States Postal Service "Express Mail Post Office to Addressee" service on the date indicated above and that this paper or fee has been addressed to Mail Stop Patent Application, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450

Anne Collette
(Typed or printed name of person mailing paper or fee)

Anne Collette
(Signature of person mailing paper or fee)

3/26/2004
(Date signed)

Field

[0001] The present invention relates to digital circuits, and more particularly, to Static Random Access Memory (SRAM) circuits.

Background

[0002] With the scaling of transistor dimensions to smaller sizes, the variability in the number and location of dopant atoms in transistor channels may result in unwanted variations in device threshold voltage among various transistors. This may be of particular concern when using minimum geometry devices in area-limited circuits, such as Static Random Access Memory (SRAM). Mismatch in threshold voltage among neighboring devices within a SRAM cell may dramatically reduce its read stability. Read cell stability may be loosely defined as the probability that during a read operation performed upon a cell, the cell will "flip" its content. This may be explained by considering Figs. 1a and 1b.

[0003] A portion of a typical SRAM is shown in simplified form in Fig. 1a, and one of its 6T SRAM cells is shown in Fig. 1b. In Fig. 1a, during a pre-charge phase, pre-charge line **102** is brought LOW so that pull-up pFETs **104** and **106** charge bitlines **108** and **110** to Vcc (HIGH). During a read operation, pre-charge line **102** is HIGH so that pull-ups **104** and **106** are OFF; and one of the wordlines, say wordline WL1, is brought HIGH so that its corresponding cell, Cell1, is read. Referring to Fig. 1b, assume that node **114** is LOW to store a logical "0" and node **116** is HIGH to store a logical "1", and that bitlines **108** and **110** have been pre-charged to Vcc. At the beginning of a read operation, wordline **112** is brought HIGH, resulting in node **114** rising above LOW ("0") due to the voltage divider comprising access nFET **118** and pull-down nFET **120**. This voltage division is between the Vcc pre-charged bitline **110** and ground node **122** (or ground rail, at voltage Vss) of the cell. If node **114** rises too high, the stored cell content may be "flipped", resulting in an incorrect read operation.

[0004] The lower the ON-resistance of nFET **120** relative to that of access nFET **118** (commonly referred to as the cell ratio), the smaller the noise figure on the "0" node (**114**). A lower noise figure, other things being equal, leads to an increase in read stability. In practice, SRAM cells should be designed to meet a specified minimum cell stability. Process scaling may make it harder to achieve this because of an expected increase in device parameter variations, e.g., variations in device threshold voltage.

[0005] Various techniques have been proposed to improve cell stability in a SRAM cell. For example, the width of the pull-down nFETs may be increased, but this results in a larger cell area and may make it more difficult to perform a stable write operation. As another example, the length of the (minimum-sized) access transistor in a SRAM cell may be increased, but this leads to a reduction in channel current during a read operation, thereby decreasing speed. As another example, the strength of the pull-down nFETs in a SRAM cell may be increased by driving their source terminals to a negative voltage just before the cell's corresponding wordline is brought HIGH. This boosts the drive of the pull-down nFETs due to increasing both the gate-to-source and drain-to-source voltages. But this requires a negative supply voltage generator with its associated area and power overhead, as well as process technology for a higher gate-oxide breakdown voltage.

Brief Description of the Drawings

[0006] Fig. 1a abstracts a prior art SRAM memory array.

[0007] Fig. 1b is a prior art SRAM memory cell at the circuit level.

[0008] Fig. 2. abstracts a SRAM memory according to an embodiment of the present invention.

[0009] Fig. 3 abstracts a body biased pFET with n-well technology.

[0010] Fig. 4 illustrates the shift in the input-output voltage transfer function of an inverter due to forward body biasing.

[0011] Fig. 5 abstracts a SRAM memory according to another embodiment of the present invention.

Description of Embodiments

[0012] An embodiment SRAM cell according to the present invention is shown in Fig. 2. For simplicity, only one 6T SRAM cell is shown having cross-coupled inverters and access transistors, where one of the inverters comprises pull-up pFET **202** and pull-down nFET **204**, the other inverter comprises pull-up pFET **206** and pull-down nFET **208**, and the access transistors comprise nFETs **210** and **212**. The bodies of pull-ups **202** and **206**, instead of being directly connected to power rail **214** at voltage Vcc as in Fig.

1b, are connected to switch **216**. Switch **216** couples the bodies of pull-ups **202** and **206** either to power rail **214** or to bias voltage generator circuit **218**. Bias voltage generator **218** provides a bias voltage Vb less than the rail voltage Vcc. When switch **216** is set to couple the bodies of pull-ups **202** and **206** to bias voltage generator **218**, pull-ups **202** and **206** are forward body biased.

[0013] Shown in Fig. 3 is a simplified, cross-sectional view of a pFET using a n-well CMOS process. Substrate **302** is a p-substrate in which n-well **304** (the "body") has been formed. Formed within n-well **304** is source/drain terminals **306**. Gate **308** is insulated from n-well **304** by insulator **310**. Fig. 3 is simplified in that not all layers are shown. (For example, for simplicity, passivation layers are not shown, nor are contacts to source/drain terminals shown.) Any suitable process technology may be utilized to form the transistor of Fig. 3. A body terminal, n^+ region **312**, is formed within n-well **304** so that the body may be biased. For simplicity, the body is shown to be biased at Vb. We may write Vb as Vb = Vcc – Vfbb, where Vfbb is a forward body bias voltage. Preferably, Vfbb should be chosen so as to prevent turning ON the parasitic junction diode formed by the source/drain terminals and the n-well. For example, in some embodiments, Vfbb may be in the range of 400 mV to 500mV.

[0014] Forward body biasing pull-ups **202** and **206** reduces their effective threshold voltage because it makes their source-to-body voltage positive. By forward body biasing pull-ups **202** and **206** during a read operation, the SRAM cell of Fig. 2 has improved read stability. To see this, assume that node **220** is LOW (stores a "0") and node **222** is HIGH (stores a "1"). With pull-up pFET **206** forward body biased, the trip point of the inverter comprising pFET **206** and nFET **208** is shifted. This shifting is illustrated in Fig. 4, where the input-output voltage transfer function of an inverter is shown being shifted to the right when forward body biasing is applied to the pFET of the inverter. As discussed with respect to Fig. 1b, during a read operation the node of **220** rises above LOW. However, with the input-output voltage transfer function of the inverter comprising pFET **206** and nFET **208** shifted to the right, node **220** needs to rise to a higher voltage to flip the cell than when pFET **206** is not forward body biased. That is, a larger noise magnitude is needed to flip the cell during a read operation. As a result, cell read stability is improved.

[0015] In other embodiments, switch **216** may not be present and the bodies of pull-ups **202** and **206** may be hardwired to bias voltage generator **218**. However, forward body biasing pull-ups **202** and **206** increases both the sub-threshold leakage current and the reverse-bias junction leakage current in these devices. Sub-threshold leakage current may be reduced by increasing the channel length of the devices in the SRAM cells. Nevertheless, in present microprocessors, embedded SRAM often makes up a sizeable portion of the total core area, so that any increase in SRAM power dissipation may be costly. Accordingly, by using switch **216**, dynamic forward body biasing may be employed to reduce power dissipation.

[0016] For example, Fig. 5 depicts, in simplified form, a large SRAM array with N·M banks, where N >> M. Upon a read or write, only M banks are accessed at a time. For example, the banks cross-hatched in Fig. 4 may be accessed during a read operation. For those M banks being accessed, their n-wells are coupled to bias voltage generator **218** so that the pull-ups in the cells of these M banks are forward body biased. All of the other banks not being accessed have their n-wells at Vcc, so that their respective pull-ups are not forward body biased. Because the bank address is known ahead of a wordline address, selected banks may be dynamically forward body biased while the wordline address is being decoded. Under this scenario, dynamic forward body biasing does not increase pipeline latency.

[0017] Switch **216** may be any device or combination of devices, for example, it may comprise a pass transistor or a transmission gate. Many types of circuits may be employed to realize bias voltage generator **218**. For example, a band-gap reference generator circuit may be used that tracks variation in Vcc, temperature, and process.

[0018] It may be noted that while forward body biasing improves read stability of a SRAM cell, it makes it harder to write. This is because a write operation begins with discharging the HIGH node of the SRAM, which is "stronger" due to the forward body biasing. As a result, it is expected that write stability may be degraded. However, in practice, read stability is often one of the main limitations in SRAM cell design, and there is usually ample margin in write stability. If, however, write stability is an issue, then a dynamic forward body biasing scenario may be employed in which forward body biasing is applied to all columns in a SRAM array except those being written to, where the ones undergoing a write operation are not forward body biased. In this case, both bank and column address should be known in advance so as not to affect latency. Furthermore, n-wells should not be shared in adjacent cells located in the same memory row.

[0019] Embodiments of the present invention may be of use in many electronic systems employing SRAM, such as the computer system illustrated in Fig. 6. In Fig. 1, die **602** comprises a microprocessor with many sub-blocks, such as arithmetic logic unit (ALU) **604** and on-die cache **606**. Die **602** may also communicate to other levels of cache, such as off-die cache **608**. Higher memory hierarchy levels, such as system memory **610**, are accessed via host bus **612** and chipset **614**. In addition, other functional units not on die **602**, such as graphics accelerator **616** and network interface controller (NIC) **618**, to name just a few, may communicate with die **602** via appropriate busses or ports. Each of these functional units may physical reside on one die or more than one die. Some or parts of more than one functional unit may reside on the same die. SRAM may used in many of these functional units. In particular, SRAM is usually used in the caches.

[0020] It is to be understood in these letters patent that the meaning of "*A* is connected to *B*" is that *A* and *B* are connected by a passive structure for making a direct electrical connection so that the voltage potential of *A* and *B* are substantially equal to each other. For example, *A* and *B* may be connected by way of an interconnect, transmission line, etc. In integrated circuit technology, the "interconnect" may be exceedingly short, comparable to the device dimension itself. For example, the gates of two transistors may be connected to each other by polysilicon or copper interconnect that is comparable to the gate length of the transistors.

[0021] It is also to be understood that the meaning of "*A* is coupled to *B*" is that either *A* and *B* are connected to each other as described above, or that, although *A* and *B* may not be connected to each other as described above, there is nevertheless a device or circuit that is connected to both *A* and *B*. This device or circuit may include active or passive circuit elements. For example, *A* may be connected to a circuit element which in turn is connected to *B*.